33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Exploring the Feasibility of Next-Generation Sequencing and Microarray Data Meta-Analysis

Po-Yen Wu, *Student Member, IEEE,* John H. Phan, *Member, IEEE,* May D. Wang, *Senior Member, IEEE*

***Abstract*—Emerging next-generation sequencing (NGS) technology potentially resolves many issues that prevent widespread clinical use of gene expression microarrays. However, the number of publicly available NGS datasets is still smaller than that of microarrays. This paper explores the possibilities for combining information from both microarray and NGS gene expression datasets for the discovery of differentially expressed genes (DEGs). We evaluate several existing methods in detecting DEGs using individual datasets as well as combined NGS and microarray datasets. Results indicate that analysis of combined NGS and microarray data is feasible, but successful detection of DEGs may depend on careful selection of algorithms as well as on data normalization and pre-processing.**

I. Introduction

MICROARRAY technology has been widely used for gene expression profiling. This technology is attractive because of its maturity and because of the large number of publicly available datasets. However, there are some inherent limitations to microarrays [1]. Starting from 2005, after 454 Life Sciences introduced its large-scale parallel pyrosequencing system, next-generation sequencing (NGS) technology gradually overtook the sequencing market. NGS is fascinating because it can identify and quantify rare transcripts without prior knowledge of a particular gene. It can also provide information regarding alternative splicing and sequence variation in identified genes [2]. Because of the distinguishing features of NGS technology such as low cost and high sensitivity compared to microarray technology, NGS is becoming the preferred method for genomic analysis.

Gene expression profiling reveals genetic mechanisms behind biological observations and facilitates biomarker-based disease diagnosis and treatment. For disease diagnosis, it is important to accurately identify DEGs between disease states. Many statistical or empirical methods have been proposed to model the behavior of the data and to identify differential expression. Each method is usually designed to work well for a particular platform. For microarray data, some common methods for identifying DEGs include Significance Analysis of Microarrays (SAM) [3], linear models and empirical Bayes [4], and Rank Products (RP) [5]. Poisson-based models—e.g., Audic-Claverie (AC), Poisson model with likelihood ratio test, and negative binomial model with exact test–are appropriate for NGS digital expression values [6-9].

Sample size has always been a problem in gene expression data analysis. Thus, meta-analysis methods—i.e., integrative methods that synthesize or review results from multiple datasets that are independent but related [10]—have been developed to alleviate biases caused by a lack of samples.

Because of the limitations of microarrays and the potential advantages of NGS, it is tempting to discard microarray samples in favor of NGS samples. However, available microarray experiments still contain valuable information. Until NGS technology matures and until samples become more widely and publically available, methods that take advantage of both the large number of microarray samples and the sensitivity of NGS samples may be useful. Existing meta-analysis methods for microarray data include Rank Products [5], mDEDS [11], GeneMeta [12], etc. In this work, we explore existing DEG detection methods, assess their performance in terms of false discovery rate, and discuss their pitfalls when combining microarray and NGS data.

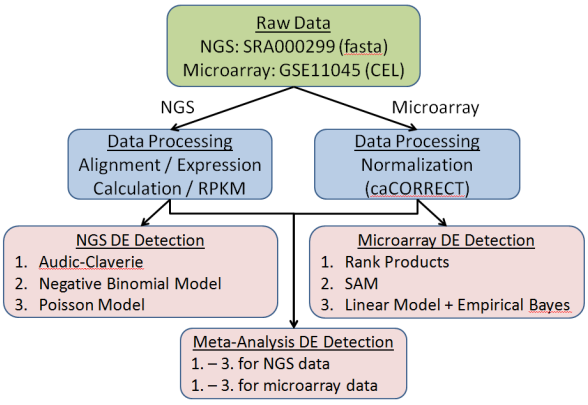

Fig. 1. Assessment of microarray and NGS meta-analysis methods involves (1) pre-processing of raw data to calculate gene expression and (2) application of several platform-specific DEG detection methods.

Manuscript received April 15, 2011. This research has been supported by grants from the Parker H. Petit Institute for Bioengineering and Bioscience (IBB), National Institutes of Health (NIH) (Bioengineering Research Partnership R01CA108468, Center for Cancer Nanotechnology Excellence U54CA119338), National Cancer Institute (NCI); Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW), and Georgia Research Alliance; Hewlett-Packard and Microsoft Research.

Po-Yen Wu is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. (e-mail: pwu33@ gatech.edu).

John H. Phan is with the Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA. (e-mail: jhphan@gatech.edu).

M. D. Wang is with the Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA. (corresponding author, phone: 404-385-2954; fax: 404-385-0383; e-mail: maywang@ bme.gatech.edu).

II. METHODOLOGY

A. Overview

We analyzed the data in three steps: (1) data acquisition, (2) data pre-processing, and (3) DEG detection (Fig. 1). The raw microarray and NGS data is available from public repositories. Both NGS and microarray data required some pre-processing to obtain gene expression values. Finally, we assessed the performance of three DEG detection methods for each platform respectively. Each method was also applied to a combined dataset of NGS and microarray samples to assess the feasibility of meta-analysis.

B. Data Acquisition

We used publicly available microarray and NGS samples from the GEO (http://www.ncbi.nlm.nih.gov/geo/) and SRA (http://www.ncbi.nlm.nih.gov/sra) databases (accession numbers: GEO11045, SRA000299) [13]. The dataset is composed of three technical replicates for each organ (kidney or liver) and platform combination. Samples were acquired using Illumina sequencing and Affymetrix GeneChip technology.

C. Microarray and NGS Data Pre-Processing

We calculated the Robust Multichip Average (RMA) gene expression values for the microarray data using the caCORRECT web-based tool [14]. We then mapped each Affymetrix probeset to an Ensembl Gene ID and discarded probesets that matched multiple Ensembl Gene IDs. This filtering is necessary for matching microarray probesets to genes in the NGS data.

We downloaded raw NGS data from the SRA database in fasta format and aligned the short sequence reads to the human genome (Genome Reference Consortium, GRCh37.61) using the bwa aligner [15] with maximum edit distance of 2 and maximum number of reported repeat alignments of 1000. If a sequence read has several possible alignment locations, it contributes a fraction to the total expression levels of the corresponding genes. For example, the expression contribution of such a sequence read is defined as $1/N$, where N is the total number of alignment locations. We then computed a standard normalized expression value: reads per kilobase of exon per million mapped reads (RPKM) of each gene that maps to an Ensembl Gene ID. Once we computed both microarray and NGS expression values, we took the intersection of both datasets to form our final datasets, which contain 16676 genes.

D. Detection of Differentially Expressed Genes

We applied three different DEG detection methods to each platform. For NGS technology, we used the (1) Audic-Claverie (AC) statistics, (2) a negative binomial model with exact test, and (3) a Poisson model with likelihood ratio test. The AC method, applicable to digital counts, calculates conditional probabilities of differential expression based on the modified Poisson distribution. Such probabilities can be treated as p-values [16]. To identify DEGs, multiple testing corrections (e.g., Benjamini & Hochberg (BH) [17]) should be performed to adjust p-values. These adjusted p-values can be treated as estimates of false discovery rate (FDR). The negative binomial model with exact test, implemented in R (package 'edgeR' [18]), uses the negative binominal distribution to model overdispersion relative to the Poisson distribution for digital gene expression data with a small number of replicates [7, 8]. It then uses an exact test to compute the exact p-values, also interpretable as FDR after BH correction. The Poisson model with likelihood ratio test has also been implemented in R (package 'DEGseq' [19]) and models RNA sequencing as a random sampling process. Each read is sampled independently and uniformly from every possible nucleotide in the sample. Thus, the number of reads coming from a gene follows a binomial distribution and can be approximated by a Poisson distribution [9]. FDR is estimated by adjusting p-values inferred from the likelihood ratio test.

For microarrays, we used (1) significance analysis of microarrays (SAM), (2) Rank Products, and (3) linear model and empirical Bayes methods. SAM has been implemented in R (package 'SAMR') and computes experimental 'relative differences' for each gene. It then uses random permutations of the data to estimate 'expected relative differences' and identifies DEGs by comparing experimental differences with expected differences. FDR is also estimated by using random permutations. Rank Products, implemented in R (package 'RankProd'), assigns a rank to each gene based on fold-changes of all inter-class pairs of samples and then calculates the rank product associated with each gene. It then uses random permutations to estimate FDR. The linear model and empirical Bayes method has also been implemented in R (package 'limma'). It fits microarray data to a linear regression model and uses an empirical Bayes method to estimate moderated t-statistics. P-values can be inferred from moderated t-statistics and can be interpreted as estimated FDR after BH multiple testing corrections.

E. Meta-Analysis

Although most DEG detection methods were designed for particular platforms, we assessed the ability of each method to handle the combined mixture of both microarray and NGS data. We considered three issues before combining datasets across platforms. First, since microarray data uses RMA normalization, which produces expression values in the $\log_2$ scale, we either transformed the NGS data into the $\log_2$ scale or transformed microarray data into the linear scale (i.e., raising to powers of 2). Second, the dynamic range of gene expression from microarray and NGS platforms are different. Thus, we used quantile normalization [20] to force data from both platforms to have identical distributions. Third, since some genes measured with NGS technology have an expression value of zero (no sequence aligned to the gene), we filtered out these genes or replaced their values (e.g., by assigning very small values) to ensure proper computation.

To test meta-analysis using each method suitable for NGS data, we used the following datasets: NGS raw data (or NGS RPKM data), microarray data that has been quantile

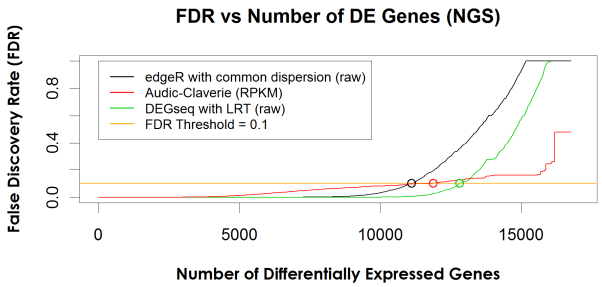

Fig. 2. False discovery rate versus the number of DEGs identified in NGS data using various NGS-specific methods.

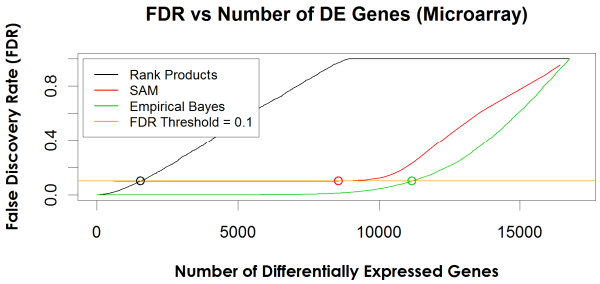

Fig. 3. False discovery rate versus the number of DEGs identified in microarray data using various microarray-specific methods.

normalized to NGS data, microarray data that has been converted to the linear scale (using power of 2 for each data point), and two possible combined datasets from previously normalized data. The AC method does not perform normalization automatically. Thus, we used RPKM expression values for the NGS data. For edgeR and DEGseq implementations, we used raw expression values of NGS data since these methods already included a normalization step.

Meta-analysis of microarray specific methods is similar. For each method, we tested the following datasets: individual microarray data, $\log_2$ transformation of individual NGS RPKM data, NGS data that is quantile normalized to microarray data, and two possible combined datasets from previously normalized data.

III. Results and Discussion

We used estimated FDR at 10% as the criterion for evaluating the performance of DEG detection methods. All genes with estimated FDR less than 10% were considered to be DEGs. We assume that more DEGs is better, since, for this study, we do not know the true set of DEGs. Moreover, we can support this assumption with the fact that biological differences are large between the two sample groups—kidney and liver tissue. If one method generally has a FDR curve lower than that of other methods, we define that method to have better performance in terms of DEG detection.

A. Performance of NGS Methods

Using three NGS-specific methods, at least 11523 DEGs were identified out of 16766 genes (Fig. 2). This means that no less than 68.7% of genes were called differentially expressed. At a FDR of 10%, the DEGseq method based on the Poisson model using the likelihood ratio test outperformed the other methods. The flat region on the right side of the AC curve is the result of low-abundance genes that have similar expression levels and thus have similar probabilities.

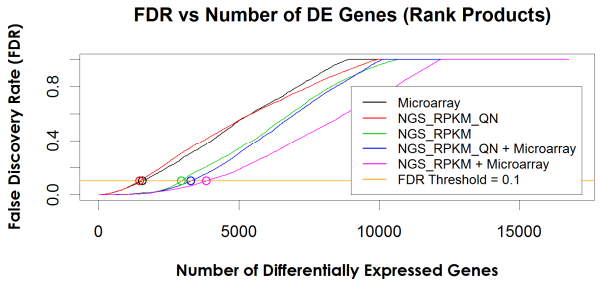

Fig. 4. False discovery rate versus the number of DEGs identified using Rank Products applied to individual NGS and microarray data as well as to combined data.

B. Performance of Microarray Methods

The number of identified DEGs ranged from 1548 (Rank Products) to 11169 (empirical Bayes) out of 16766 genes (Fig. 3). Empirical Bayes produced the best result with 66.6% of genes called differentially expressed.

The large differences in FDR curves may result from the design of FDR estimation algorithms of these methods. SAM and Rank Products estimate FDR by random permutation, and then set thresholds to count the number of false positives. Rank Products permutes values within samples, allowing a much larger number of permutations. On the other hand, SAM permutes sample tags (class labels) and may not work well when sample size is small.

C. Meta-Analysis of NGS and Microarray Data

Table I summarizes all of the meta-analysis results. We tested the performance of each method on five datasets, representing different combinations of normalized microarray and NGS data. We observed that AC statistics, DEGseq, and Rank Products are able to detect more DEGs when microarray and NGS data are combined. This suggests that these methods may be more robust to heterogeneous combination of microarray and NGS data. The other three

TABLE I
NUMBER OF DEGS @ FDR = 10% (META-ANALYSIS)

Dataset	Audic-Claverie	edgeR	DEGseq (LRT)
NGS raw (or RPKM) data	11875	**11491**	13138
Microarray QN to NGS data	8372	10276	10877
Microarray power of 2 data	N.A.	8018	8276
NGS + Microarray (QN)	**13903**	2185	**13289**
NGS + Microarray (power)	N.A.	1913	12350

Dataset	SAM	Rank Products	Empirical Bayes
Microarray data	8559	1552	11169
NGS QN to microarray data	6222	1456	**11556**
$\log_2$ NGS RPKM data	**9121**	2950	11383
Microarray + NGS (QN)	2410	3290	4809
Microarray + NGS ($\log_2$)	997	**3836**	80

* QN: quantile normalization; LRT: likelihood ration test; power: transform from log to linear scale.
* N.A. in AC statistics means that this value is not trustworthy, reporting an unreasonable DEG detection rate of 100%.

methods, edgeR, SAM, and Empirical Bayes did not perform well when combining datasets compared to individual datasets. However, more datasets should be tested before drawing conclusions.

We also evaluated the behavior of FDR versus the number of identified DEGs for the Rank Products method (Fig. 4). The FDR at multiple thresholds (number of genes selected as DEGs) for the Rank Products method decreases when sample size increases (by combining NGS and microarray datasets). In Fig. 4, the combined dataset curves are represented by the blue and magenta lines. But NGS data did not benefit from quantile normalization, the FDR increases in this case (Fig .4, red line and blue line).

Some meta-analysis methods (e.g., SAM) results in an increased FDR compared to that of individual data, regardless of the normalization method. These methods depend on assumptions about expression values (i.e., expression values for each sample are assumed to be drawn from particular distributions). Statistically, the dynamic range and data distributions differ between microarray and NGS platforms. The use of normalization methods such as RPKM and quantile normalization is sometimes not enough to overcome these differences. Methods such as Rank Products are designed for meta-analysis and are less stringent in terms of distribution similarity between datasets.

IV. Conclusion

We have assessed six methods for detecting DEGs from NGS and microarray data. Results indicate that these methods, applied to individual datasets, vary widely in terms of number of DEGs detected at fixed FDR. We have also evaluated the ability of each method for meta-analysis of NGS and microarray data (combined analysis). We determined that such analysis is possible, but should be applied cautiously. Normalization and data transformation methods can affect the performance of these methods. Using the limited datasets in this study, some methods appear to benefit from normalization and transformation, for example, Rank Products, DEGseq, and AC statistics. In contrast, some methods do not perform well when data is normalized, for example, SAM, edgeR, and empirical Bayes. Thus, results of our exploratory analysis indicate that it is feasible to use meta-analysis methods for microarray and NGS data. Such analysis takes advantage of both the large number of available microarray samples and the higher sensitivity of NGS samples for detecting DEGs. However, the success of meta-analysis may also depend on normalization or other pre-processing steps.

References

[1] B. R. Graveley, "Molecular biology: power sequencing," *Nature,* vol. 453, pp. 1197-8, Jun 26 2008.

[2] M. L. Metzker, "Sequencing technologies - the next generation," *Nat Rev Genet,* vol. 11, pp. 31-46, Jan 2010.

[3] V. G. Tusher, *et al.*, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A,* vol. 98, pp. 5116-21, Apr 24 2001.

[4] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Stat Appl Genet Mol Biol,* vol. 3, p. Article3, 2004.

[5] R. Breitling, *et al.*, "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments," *FEBS Lett,* vol. 573, pp. 83-92, Aug 27 2004.

[6] S. Audic and J. M. Claverie, "The significance of digital gene expression profiles," *Genome Res,* vol. 7, pp. 986-95, Oct 1997.

[7] M. D. Robinson and G. K. Smyth, "Moderated statistical tests for assessing differences in tag abundance," *Bioinformatics,* vol. 23, pp. 2881-7, Nov 1 2007.

[8] M. D. Robinson and G. K. Smyth, "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," *Biostatistics,* vol. 9, pp. 321-32, Apr 2008.

[9] H. Jiang and W. H. Wong, "Statistical inferences for isoform expression in RNA-Seq," *Bioinformatics,* vol. 25, pp. 1026-32, Apr 15 2009.

[10] S. L. Normand, "Meta-analysis: formulating, evaluating, combining, and reporting," *Stat Med,* vol. 18, pp. 321-59, Feb 15 1999.

[11] A. Campain and Y. H. Yang, "Comparison study of microarray meta-analysis methods," *BMC Bioinformatics,* vol. 11, p. 408, 2010.

[12] J. K. Choi, *et al.*, "Combining multiple microarray studies and modeling interstudy variation," *Bioinformatics,* vol. 19 Suppl 1, pp. i84-90, 2003.

[13] J. C. Marioni, *et al.*, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Res,* vol. 18, pp. 1509-17, Sep 2008.

[14] T. H. Stokes, *et al.*, "chip artifact CORRECTion (caCORRECT): a bioinformatics system for quality assurance of genomics and proteomics array data," *Ann Biomed Eng,* vol. 35, pp. 1068-80, Jun 2007.

[15] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics,* vol. 25, pp. 1754-60, Jul 15 2009.

[16] J. R. Bradford, *et al.*, "A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling," *BMC Genomics,* vol. 11, p. 282, 2010.

[17] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society Series B-Methodological,* vol. 57, pp. 289-300, 1995.

[18] M. D. Robinson, *et al.*, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics,* vol. 26, pp. 139-40, Jan 1 2010.

[19] L. Wang, *et al.*, "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data," *Bioinformatics,* vol. 26, pp. 136-8, Jan 1 2010.

[20] B. M. Bolstad, *et al.*, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics,* vol. 19, pp. 185-93, Jan 22 2003.